Data Quality Plan

Tackling data quality issues identified in the data quality report

Issues rectified before the quality report:

1.1 Converting the data types

The first decision to make was identifying which features should be categorised in what way. Three features were selected as "categorical" as they represented a finite number of possible states. These features were:

- RiskPerformance (evaluated to good or bad)
- MaxDelq2PublicRecLast12M (range from 1-7)
- MaxDelqEver (range 2-8)

Now it was necessary to choose which columns would be represented by continuous columns. I chose the following features to be considered as continuous columns:

- ExternalRiskEstimate
- PercentTradesNeverDelq
- PercentInstallTrades
- NetFractionRevolvingBurden
- NetFractionInstallBurden
- NumRevolvingTradeWBalance
- NumInstallTradeWBalance
- PercentTradesWBalance
- AverageMInFile
- MSinceOldestTradeOpen
- NumSatisfactoryTrades
- MSinceMostRecentTradeOpen
- NumTrades60Ever2DerogPubRec
- NumTrades90Ever2DerogPubRec
- MSinceMostRecentDelq
- NumTotalTrades
- NumTradesOpenInLast12M
- MSinceMostRecentInqexcl17days
- NumInqLast6M
- NumInqLast6Mexcl17days
- NumBank2NatlTradesWHighUtilization

1.2 Eliminating duplicate rows

The next step was to eliminate the duplicate rows. There were many rows containing the -9 value. As per the context dictionary, this holds a special meaning “no bureau record or investigation”. However, considering the -9 value spanned the entire row, there was very little useful information (except, maybe, the number of accounts that were never investigated and still had a credit rating) to correlate alongside the good or bad rating. Therefore, I justifiably dropped them from the data frame.

Issues tackled after quality report complete:

Next Page

Feature	Data Quality Issue/s	Handling Strategy
Multiple Features	-9 special meaning combined with	-9 Rows are not correlated with any useful data in their respective instances. Therefore, I believe they can be safely dropped.
Multiple Features	-8 Values for columns with low incidence of this.	For features were this isn't a regular occurrence, we can impute the mean as it shouldn't affect the variance significantly.
RiskPerformance	Missing Values	Complete Case Analysis (remove missing values if the value is a target feature)
MSinceMostRecentDelq	Close to 50% of data in this feature is -7, meaning the condition is not met. Presence of -8 special meaning values	Set these instances to NaN as a flag indicator for "criteria not satisfied". This allows us to do more investigation between features where these values are present/absent later. Replace –8 Values with the mean() for the feature to avoid mixing -7 and -8 both in NaN
PercentTradesNeverDelq	~26% of accounts with having "never had delinquency" but not recorded has "current and never delinquent" in MaxDelqEver.	Convert these values NaN (missing) as this appears to be erroneous data.
NetFractionRevolvingBurden	Small percentage of missing data (-8 special value)	Impute the mean of this feature
NetFractionInstallBurden	Presence of 37% missing values (-8 special meaning).	Set these values to NaN as a flag indicator, so we can still use remaining data to look for correlation between features (where this value was present and where it wasn't).
NumRevolvingTradesWBalance	Small percentage of missing data (-8 special value) Outliers	Impute the mean of this feature Do nothing

NumInstallTradesWBalance	Small percentage of missing data (-8 special value) Outliers	Impute the mean of this feature Do nothing
NumBank2NatlTradesWHighUtilization	Small percentage of missing data (-8 special value) Outliers	Impute the mean of this feature Do nothing
PercentTradesWBalance	Small percentage of missing data (-8 special value) Outliers	Impute the mean of this feature Do nothing
MSinceOldestTradeOpen	Small percentage of missing data (-8 special value) Outliers	Impute the mean of this feature Do nothing
MSinceMostRecentTradeOpen	Outliers	Do nothing
AverageMInFile	Outliers	Do nothing
NumSatisfactoryTrades	Outliers	Do nothing
NumTrades60Ever2DerogPubRec	Outliers	Do nothing
NumTrades90Ever2DerogPubRec	Outliers	Do nothing
NumTotalTrades	Outliers	Do nothing
NumTradesOpeninLast12M	Outliers	Do nothing
NumInqLast6Mexcl7days	Outliers	Do nothing

Cleaned data results

2.1 Continuous features

	count	mean	std	min	25%	50%	75%	max
ExternalRiskEstimate	898.0	71.922049	9.705045	40.0	65.00	72.0	79.750000	93.0
MSinceOldestTradeOpen	898.0	199.157714	93.069484	23.0	135.25	185.0	261.000000	530.0
MSinceMostRecentTradeOpen	898.0	9.497773	11.938457	0.0	3.00	6.0	12.000000	156.0
AverageMInFile	898.0	77.979955	32.744401	9.0	57.00	75.0	95.000000	257.0
NumSatisfactoryTrades	898.0	20.582405	11.160649	1.0	12.00	19.0	27.000000	63.0
NumTrades60Ever2DerogPubRec	898.0	0.723831	1.714100	0.0	0.00	0.0	1.000000	19.0
NumTrades90Ever2DerogPubRec	898.0	0.489978	1.499316	0.0	0.00	0.0	0.000000	19.0
PercentTradesNeverDelq	684.0	89.611111	12.451167	27.0	86.00	94.0	100.000000	100.0
MSinceMostRecentDelq	481.0	21.536102	19.960770	0.0	5.00	16.0	31.000000	82.0
NumTotalTrades	898.0	21.717149	12.302403	0.0	13.00	20.0	29.000000	68.0
NumTradesOpeninLast12M	898.0	1.914254	1.890656	0.0	0.00	1.0	3.000000	12.0
PercentInstallTrades	898.0	35.631403	17.242613	0.0	24.00	33.0	46.000000	100.0
MSinceMostRecentInqexcl7days	689.0	2.399129	4.617528	0.0	0.00	0.0	3.000000	23.0
NumInqLast6M	898.0	1.456570	1.966499	0.0	0.00	1.0	2.000000	21.0
NumInqLast6Mexcl7days	898.0	1.390869	1.937083	0.0	0.00	1.0	2.000000	21.0
NetFractionRevolvingBurden	898.0	35.832579	29.402319	0.0	9.00	32.0	57.000000	165.0
NetFractionInstallBurden	577.0	69.183709	25.083836	2.0	54.00	75.0	88.000000	140.0
NumRevolvingTradesWBalance	898.0	3.913093	2.939052	0.0	2.00	3.0	5.000000	23.0
NumInstallTradesWBalance	898.0	2.544567	1.600976	1.0	2.00	2.0	3.000000	19.0
NumBank2NatlTradesWHighUtilization	898.0	1.086698	1.509048	0.0	0.00	1.0	1.086698	16.0
PercentTradesWBalance	898.0	66.872910	22.566920	0.0	50.00	67.0	83.000000	100.0

2.2 Categorical features

	count	unique	top	freq
RiskPerformance	898	2	Bad	472
MaxDelq2PublicRecLast12M	898	9	7	382
MaxDelqEver	898	7	8	408